Item Position and Item Difficulty Change in an IRT-Based Common Item Equating Design

Jason L. Meyers
Pearson

G. Edward Miller
Texas Education Agency

Walter D. Way
Pearson

In operational testing programs using item response theory (IRT), item parameter invariance is threatened when an item appears in a different location on the live test than it did when it was field tested. This study utilizes data from a large state’s assessments to model change in Rasch item difficulty (RID) as a function of item position change, test level, test content, and item format. As a follow-up to the real data analysis, a simulation study was performed to assess the effect of item position change on equating. Results from this study indicate that item position change significantly affects change in RID. In addition, although the test construction procedures used in the investigated state seem to somewhat mitigate the impact of item position change, equating results might be impacted in testing programs where other test construction practices or equating methods are utilized.

INTRODUCTION

Item response theory (IRT) applications typically utilize the characteristic of item parameter invariance. That is, when the underlying assumptions are satisfied, IRT supports the practice of applying item parameter estimates obtained from a
previous sample of test-takers to new samples. In theory, the advantages of item parameter invariance permit applications such as computerized-adaptive testing (CAT) and test pre-equating.

A number of threats to item parameter invariance exist at the item level. These include context effects, item position effects, instructional effects, variable sample sizes, and other sources of item parameter drift that are not formally recognized or controlled for in IRT applications. Several researchers have documented the existence and investigated the influence of such item level effects (Whitley & Dawis, 1976; Yen, 1980; Klein & Bolus, 1983; Kingston & Dorans, 1984; Rubin & Mott, 1984; Leary & Dorans, 1985).

In operational testing programs using IRT, model advantages must often be weighed against concerns over threats to item parameter invariance. One place where this tension occurs is the position of items in a test from one use to the next. The conservative view is that item sets used for test equating should be kept in identical or very similar positions within the test from one use to the next. Of course, this requirement can be limiting and difficult to sustain over time. For example, using items more than once in operational administrations may be a concern, either for security reasons and/or because of test disclosure requirements. In addition, test content and design considerations often require at least some shifting in the location of common items across test administrations. Many state departments of education face both of the above. It is common for states to be required to release their operational test items after test administration. This makes embedding field test items on multiple forms necessary. It is virtually impossible to construct new tests from these sets of embedded field test items with the item positions intact. In addition, reading tests contain items that are connected to a reading passage and, therefore, must stay together. This adds to the problem of states being able to maintain item position from field test to operational test.

The most liberal stance with respect to position effects is implicit in CAT, in which items are administered by design in varied positions that are difficult to predict. CAT assumes that item position effects may be ignored. In practice, it is difficult to investigate this assumption because CAT data are sparse and because each test-taker typically receives a unique test.

Most scaling and equating designs fall between the extremes of re-administering linking items in exactly the same positions and the typical CAT administration. In some IRT common item equating designs, item parameter estimates from several previously administered forms may be used to link a new operational test form to a common scale. Two procedures are commonly used in such designs to mitigate the potential impact of position and other context effects. First, item parameters are re-estimated based on the operational data and linked back to the previous estimates (as opposed to developing pre-equated conversions based on the initial estimates). Second, a screening process is typically used to eliminate items from the common item set if the new and previous estimates differ by more than some
threshold (Miller, Rotou, & Twing, 2004). There is disagreement among psychometricians over what this threshold should be. Theoretically, the threshold should be related to sample size. Wright and Douglas (1975, pp. 35–39), however, note that random uncertainty in item difficulty of less than 0.3 logits has no practical bearing on person measurement. As a consequence, a number of Rasch-based assessment programs use the threshold of 0.3.

Because operational testing programs often require scaling and equating approaches where the positions of linking items vary across administrations, it is important to examine position effects and to try to quantify these effects in relation to other possible sources of variation in item performance. The purpose of this study was to analyze and model the effects of item position on item difficulty estimates in a large-scale K–12 testing program. The testing program utilizes a Rasch-based, common item equating design where the operational tests are equated by re-estimating item parameters and linking them back to field-test estimates. Because the field-testing occurs by embedding field-test items within the operational test, the positions of the items within the test change between field-testing and operational use. Changes in Rasch item difficulties (RIDs) from field-testing to operational testing were modeled as a function of changes in item position, test level, test content, and item format (e.g., passage-based vs. discrete). Based on the final models, a simulation study was also performed to assess the effect of item position change on equating.

RELATED LITERATURE

A number of researchers have investigated the effect of item position (or the effect of changing the item position) on item difficulty. Tests varying across age levels and purposes have been considered in these studies: from elementary grade reading and math tests to post-graduate level bar and licensure exams. Reviewed in this article are studies published post-1975; an excellent review of studies performed prior to 1975 (with a few later than 1975) is provided by Leary and Dorans (1982).

Many studies have documented the finding that changing the position of items in a test can affect item difficulty. Whitely and Dawis (1976) found significantly different Rasch item difficulty estimates for 6 of 15 core items that differed only in item position across forms. Eignor and Stocking (1986) also found that the positioning of SAT items from pretest to operational test appeared to make a difference in the item difficulty. Harris (1991) analyzed three forms of the ACT with items located in different positions and found that different scale scores would have been obtained for some students. The PISA (Programme for International Student Assessment) 2000 Technical Report identified a booklet (or item position) effect with the only difference in the different
test booklets being item arrangement. Haertel (2004) found that differences in linking item position between the year 1 and year 2 tests caused anomalous behavior in linking items. Finally, Davis and Ferdous (2005) identified item position effects on item difficulty for a grade 3 reading test, and for math and reading tests at grade 5. From these studies it is fairly evident that when items change positions between testing administrations, the associated item difficulties can change as well.

In addition to impacting item difficulty, change in item position has been shown to affect equating results. Yen (1980) investigated item position on the California Achievement Test (CAT) and found some effects on parameter estimates as well as an impact on equating results. Similarly, Kingston and Dorans (1982) found that item position effects had an adverse effect on equating forms of the Graduate Record Examination (GRE). Brennan (1992) found that equating of the ACT was affected by the scrambling of items on different forms. Kolen and Harris (1990) found that when ACT mathematics items were pretested in a separate section (presumably at the end of the test), effects due to a lack of motivation or fatigue had a subsequent adverse impact on equating results. Zwick (1991) attributed a scaling anomaly between the 1984 and 1986 reading NAEP to differences in item positions. Pommerich and Harris (2003) identified an item position effect for both ACT Math and Reading and concluded that different equating relations may be obtained as a result of different item positions. In each of these studies, item position changes led not only to changes in individual item difficulties, but also contributed to differences in test equating results.

Other studies have identified more specific impacts of changing item position. Wise, Chia, and Park (1989) found that item order and other item context alterations impacted low-achieving test-takers more than high-achieving test-takers on tests of work knowledge and arithmetic reasoning. In addition, Way, Carey, and Golub-Smith (1992) found that TOEFL reading comprehension items that were pretested toward the beginning of the section but that appeared toward the end of the section on the operational test became more difficult, but items pretested near the end of the section and operationally tested near the beginning of the section became easier. These studies suggest that item position effects can vary depending on factors such as how much the item position changes, the direction of change (i.e., toward the beginning of the test versus toward the end of the test), and the ability levels of the test-takers.

Taken as a whole, these studies consistently indicate effects on IRT-based item parameter estimates when items change positions between administrations. In this study, our interest was in modeling these effects based on data from a specific testing program, and in investigating the implications of these effects for the equating procedures used to support the program.
MODELING OF CHANGE IN ITEM DIFFICULTY FROM FIELD TEST TO OPERATIONAL TEST

The data used for modeling the change in Rasch item difficulty (RID) from field-testing to operational testing came from an assessment program in a large state in which students take standardized tests in the areas of reading, mathematics, writing, social studies, and science. These assessments have been administered since the 2003 school year. In grades 3 through 8, the tests contain between 40 and 50 multiple choice items. They are untimed, and students have up to an entire school day to complete the assessment. In fact, there have been reports of students staying well into the evening hours in order to finish testing. The assessments in this state are independently scaled for each grade and subject using the Rasch Model. The item statistics are those of items that appeared on live tests in spring 2004. The statistics are based on exceptionally large sample sizes, with live test items being taken by several hundred thousand students.

An analysis of item position change effects was particularly important for this testing program because the field-tested item difficulties are utilized in the operational form equating procedures. For this testing program, the need to disclose complete operational test forms makes it infeasible to maintain an internal anchor set of equating items from one test administration to the next. As a result, a field-test design is employed in which the operational item parameter estimates are scaled through the parameter estimates obtained when the items were field-testing. In this design, all of the operational items are effectively the common item anchor set. However, the design requires changing the positions of the items between field testing and use in the final forms.

Only data from reading and mathematics were analyzed. In addition, only grades 3 through 8 were used, because these grades are based on a common objective structure; the curriculum shifts substantially at grade 9. Descriptive statistics for the items appearing at each grade and subject are presented in Table 1.

For each grade and subject, the table displays the number of items on each test, the average field test Rasch item difficulty and its associated standard error, minimum and maximum value, and the average live test RID and its associated standard error, minimum and maximum value. On average, the field tests and live tests were of roughly equal difficulty.

In addition, Table 2 displays—for all grades and subjects—the average RID change between field test and live test based on the number of item positions an item moved. It should be noted that for a given grade and subject, the field test and live test contain the same number of items. So, for example, an item appearing as item number 7 on the field test and item number 7 on the live test would have an item position change of zero. As shown in the table, the majority of items moved at least 5 positions in either direction, and many moved more than 20 positions.
As expected, the average RID change decreases as item position decreases and increases as item position increases.

Math Analyses

Using multiple regression, change in RID was first modeled as a function of item position change (item position), grade, objective, and time between field testing.
and live testing \((time)\). Because each grade and subject is scaled independently, pre- and posttest RIDs were standardized to unit normal; this allowed for meaningful comparisons across grade and subject. Regressions were conducted separately for each grade, and then across grades using grade as a predictor. This resulted in a total of seven regressions.

Inspection of the parameter estimates and \(t\)-test \(p\)-values for all seven regressions suggested that all independent variables other than \(item\ position\) and \(time\) could be dropped from the analyses. Full versus Reduced model tests were conducted to confirm this. Because \(grade\) was dropped from the combined 3–8 model, no further analysis of the separate single grade datasets was deemed appropriate.

At this point, only \(item\ position\) and \(time\) remained as predictors in the analysis. However, a closer look at the original regression output revealed that the variable \(time\) should also be dropped. \(Time\) was only significant for grade 6, it had a negative coefficient (which suggests that over time the items become more difficult—contrary to research), and only one item was making the coefficient statistically significant. \(Time\) was dropped from subsequent models and it was decided that only items field tested one year prior to live testing would be included in future analyses.

The mathematics test model now included only the single explanatory variable \(item\ position\). A scatterplot of \(item\ position\) by change in RID (Figure 1) indicated a curvilinear relationship. Both quadratic and cubic models were estimated; the cubic model provided the closest fit. Even so, the cubic model predicted close to a 0.06 RID change with no item position change. Knowing from the actual data that the mean RID difference for item position differences of 5 or less was approximately 0.005, the cubic model was re-estimated with the intercept constrained to be zero. The final mathematics tests model was

\[
Y = 0.00329x + 0.00002173x^2 + 0.00000677x^3
\]

\(\text{FIGURE 1 RID difference by item position change for grades 3–8 math combined.}\)
The mathematics regression model had an R-square value of 0.5604, indicating that about 56% of the variance in change in Rasch item difficulty could be attributed to item position change. One may note that the model indicates that the effect of a change in item position is asymmetric about zero; specifically the model indicates that placing an item nearer the end of the test has slightly more effect on its item difficulty than placing it nearer the beginning of the test. Yen (1980) and Wise et al. (1989) found that fatigue impacts low-achieving test-takers more than high-achieving test-takers and that this effect is most pronounced toward the end of the test. Thus, the asymmetry of the regression equation produced appears to be supported by previous research. It should also be noted that a cross-validation of the regression equations was considered for this study, but was ultimately decided against due to the large sample sizes and the fact that the regression equations were only used for demonstration purposes.

Reading Analyses

Change in RID for the reading tests was modeled as a function of item position, grade, and time. Objective was excluded due to (a) possible objective-passage confounding and (b) objective-passage combination sample size considerations. However, the consideration that change in RID might be affected by the reading passage that the item is associated with had to be taken into account. Consequently, the reading test data were initially modeled using a two-level random coefficient model where items were nested within passages. Separate random coefficients regressions by grade were performed along with a single random coefficients regression on the combined-grades dataset.

Results from these regressions indicated that passage variability was not statistically significant and, thus, standard ordinary least squares (OLS) regression was used for all the subsequent analyses.

As was the case in mathematics, the variable grade (in the combined data regression analysis) was not found to be statistically significant in any of the regressions. Additionally, as observed in all but grade 6 of the mathematics tests, the variable time was not statistically significant for any reading tests. Thus, all further analysis of reading data was restricted to the combined grades data with the single independent variable item position.

A scatterplot of item position by change in RID (Figure 2) indicated that the relationship between the two variables was curvilinear. Both quadratic and cubic
models were estimated; the cubic model provided the closest fit. Finally, as had been done with math, the cubic model was re-estimated with the intercept constrained to be zero. The final reading tests model was

\[
\text{rid}_{\text{difference, READ}} = 0.00845 (\text{item position}) - 0.00008343 (\text{item position})^2 + 0.00001135 (\text{item position})^3
\]

The reading regression model had an R-square value of 0.7326, indicating that about 73% of the variance in change in Rasch item difficulty could be attributed to item position change.

**SIMULATION STUDY-METHOD**

To further illustrate the effects of item position change in the context of the test equating procedures used in this testing program, demonstration simulations were conducted using a set of 281 field-test items from the grade 5 mathematics test. We chose not to simulate reading data because the modeling results for reading and math were similar and a single set of simulations seemed sufficient to demonstrate the effects in question. The mean, standard deviation, minimum, and maximum of the RIDs for the mathematics items were 0.20, 1.32, –3.43, and 3.21, respectively. These items were spiraled in sets of eight in different operational test form versions containing 44 items. The field-test item difficulty estimates were used as true
item parameters for the simulations, which repeated the following steps for each of 100 iterations:

1. Create a set of “field-test” item parameter estimates by “perturbing” (i.e. adding an error component to) the true parameters, with error components based on $SE_{bi}$ assuming a particular distribution of ability ($\theta$) and sample size.

   First, for the Rasch item difficulty (RID) of each field test item $i$, $i = 1, \ldots, 281$, compute its information as follows:

   $$ I(b_i) = \sum_{j=0}^{44} P_{ij} (1 - P_{ij}) \pi_{\theta_j} N_{FT} $$

   where

   $$ P_{ij} = [1 + e^{b_i - \theta_j}]^{-1} $$

   $b_i$ = “true” RID for item $i$

   $\pi_{\theta_j}$ = assumed proportion of students with ability $\theta_i$

   $N_{FT}$ = assumed number of students field tested

   The standard error for each $b_i$ is then computed as follows:

   $$ SE_{b_i} = \frac{1}{\sqrt{I(b_i)}} $$

   Finally, the perturbed “field test” item parameter estimate for each item is

   $$ \hat{b}_i = b_i + z_{random}(SE_{b_i}) $$

   where $z_{random}$ = random variate from $N(0,1)$ distribution.

2. Construct a test by selecting at random from the available pool of 281 field-test items a fixed number of items across various difficulty strata that were established from analyzing a final test form used operationally. The selected items were then ordered by difficulty in a manner that modeled operational practice. Figure 3 presents the field-test Rasch difficulty values plotted against their operational item positions for the mathematics tests at grades 3 through 8. As can be seen in Figure 3, these tests were constructed by placing easier items toward the beginning and end of the test, and more difficult items in the middle of the test. (Note that new field-test items are placed in positions 21 to 30.) The relationship between Rasch difficulty and item position is fit reasonably well by a polynomial regression. This relationship was used to order the items in the simulations.
3. For the items selected, create a set of “final-form” item parameter estimates by perturbing the true parameter components based on SE_b as in step 1, assuming a sample size that was five times the corresponding field-test sample size. In addition, a fixed component was added to the true item difficulty that was calculated by entering the difference between each item’s position field-tested and final form item positions into Equation 1.

For the RID of each field test item $i, i = 1, \ldots, 281$, compute its information as follows:

$$I(b_i) = \sum_{j=0}^{44} P_{ij} (1 - P_{ij}) \pi_{\theta_j} N_{FF}$$

where

$$P_{ij} = \left[1 + e^{\theta_j - b_i}\right]^{-1}$$
ITEM POSITION CHANGE

49

bi = “true” RID for item i
πθi = assumed proportion of students with ability θi
N_FF = assumed number of students tested with “Final Form”

The standard error for each bi is then:

\[ SE_{b_i} = \frac{1}{\sqrt{I(b_i)}} \]

The perturbed “final form” item parameter estimate for each item is

\[ \hat{b}_i = b_i + (0.00329\Delta + 0.00002173\Delta^2 + 0.0000067723\Delta^3) + z_{\text{random}}(SE_{b_i}) \]

where \( \Delta = \text{final form item position - field test item position} \)
\( z_{\text{random}} = \text{random variate from N}(0,1) \) distribution.

4. Link the final form item parameter estimates to the field-test estimates by calculating the mean difference between the two sets of estimates. As part of the linking, an iterative “stability check” was utilized, through which items with differences between field-test and final form difficulty estimates greater than 0.3 were eliminated from the linking item set. This step imitated operational procedures used in the testing program, which are commonly employed in IRT test equating settings.

5. Generate a test characteristic curve (TCC) based on the final linked item parameter estimates and compare this to the true TCC based on the true item parameters.

The outcomes of interest over the 100 replications included the true minus estimated TCC differences (evaluated over fixed theta levels), the differences between the mean of the true and final estimated item difficulties (i.e., the equating constant), and the number of items removed in the iterative stability checks across the 100 replications. To vary the values of SEb, four variations of field-test and final sample sizes were used: 500 and 2,500, 1,000 and 5,000, 2,000 and 10,000, and 20,000 and 100,000. For each sample size combination, one set of 100 replications simulated the systematic error due to the calculated changes in item position and a second set of 100 replications did not incorporate this systematic error. In the second set of 100 replications, step 3 as described earlier included the \( z_{\text{random}}(SE_{bi}) \) component but did not include the component based on the function of item position change. This resulted in eight sets of replicated simulations in total.
SIMULATION STUDY-RESULTS

Figure 4 presents the relationships between the field-test Rasch difficulties and the final form item positions for the test forms selected in 10 simulation iterations of the grade 5 math tests. It can be seen from comparing Figure 3 and Figure 4 that the relationships between item difficulty and item position in the simulations are quite similar to the relationships seen in the operational tests.

Table 3 summarizes the results from the various simulation study conditions. As shown in the table, the number of items remaining in the equating set after the iterative stability check removed items with RID changes of 0.30 or greater from field test to final form increased as the sample sizes increased. Fewer items remained in the equating set in the conditions where the effect of item position change was incorporated into the model, although once the sample size reached

\[ y = -0.0024x^2 + 0.1269x - 1.036 \]
\[ R^2 = 0.3336 \]
ITEM POSITION CHANGE

2,000 on the field test and 10,000 on the final form there were no observable differences.

Table 3 also shows that the magnitude of the equating constant decreased as sample size increased. Again, in the conditions in which the effect of item position change was incorporated, the equating constants were larger than when it was not included. However, once the sample size reached 2,000 on the field test, the impact of item position change was negligible.

Finally, the mean differences between the true and estimated test characteristic curves are shown in Table 3 and revealed a similar pattern. The larger the sample size, the more accurate were the estimated TCCs. Item position change impacted the accuracy slightly, although once the sample size reached 2,000 on the field test the differences between the conditions simulating the impact of item position change and those ignoring it were minimal.

Figure 5 summarizes the true minus estimated TCC differences over fixed theta levels for simulation sample sizes of 500 (field-test) and 2,500 (final form). The upper panel shows results where no item position effects were simulated and the lower panel shows results based on incorporating the item position effects into the simulations. In these graphs, the plotted points represent the mean difference between the true and estimated TCC equivalents over the 100 simulated replications at each true theta level, and the vertical lines extending from the points are based on one standard deviation of the differences between the true and estimated TCC equivalents over the 100 replications. These results indicate only a very small effect due to changes in item position. The reason for this finding is the practice followed for ordering the items in the final operational forms. Because easy and difficult items are placed in the operational forms symmetrically around the positions where the items were originally field-tested, the effects of item position changes

### TABLE 3
Summary of Grade 5 Math Simulation Results

<table>
<thead>
<tr>
<th>Condition</th>
<th>Items Remaining</th>
<th>Equating Constant</th>
<th>True-Estimated TCCs</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>M(SD) Min,Max</td>
<td>M(SD) Min,Max</td>
<td>M(SD) Min,Max</td>
</tr>
<tr>
<td>500/2500*</td>
<td>39.51(2.11) (35,44)</td>
<td>-0.016(0.033) (-0.100,0.080)</td>
<td>-0.007(0.009) (-0.020,0.003)</td>
</tr>
<tr>
<td>500/2500</td>
<td>41.96(1.21) (39,44)</td>
<td>-0.001(0.024) (-0.056,0.055)</td>
<td>-0.005(0.007) (-0.014,0.005)</td>
</tr>
<tr>
<td>1000/5000*</td>
<td>41.78(1.14) (38,44)</td>
<td>-0.012(0.019) (0.054,0.050)</td>
<td>-0.009(0.005) (-0.016,0.000)</td>
</tr>
<tr>
<td>1000/5000</td>
<td>43.65(0.58) (42,44)</td>
<td>-0.001(0.017) (-0.049,0.045)</td>
<td>-0.014(0.008) (-0.023,0.000)</td>
</tr>
<tr>
<td>2000/10000*</td>
<td>44(0) (44,44)</td>
<td>-0.000 (0.011) (-0.033,0.029)</td>
<td>-0.007(0.004) (-0.013,0.000)</td>
</tr>
<tr>
<td>2000/10000</td>
<td>44(0) (44,44)</td>
<td>-0.001(0.011) (-0.029,0.027)</td>
<td>-0.006(0.003) (-0.009,0.000)</td>
</tr>
<tr>
<td>20000/100000*</td>
<td>44(0) (44,44)</td>
<td>-0.000 (0.004) (-0.013,0.012)</td>
<td>-0.004(0.002) (-0.006,0.000)</td>
</tr>
<tr>
<td>20000/100000</td>
<td>44(0) (44,44)</td>
<td>-0.000(0.004) (-0.009,0.009)</td>
<td>-0.003(0.002) (-0.004,0.000)</td>
</tr>
</tbody>
</table>

Note. * = effect of item position change simulated.
largely cancel out at the level of the test characteristic curve. Figures 6 through 8 present the corresponding TCCs for the 1,000/5,000, 2,000/10,000 and 5,000/100,000 conditions. These figures indicate that effects due to changes in item position decreased even more as sample sizes increased.

Sequencing a test form so that the most difficult items appear in the middle of the test is somewhat unusual, as conventional test development practices typically order items so that the easier items appear at the beginning of the test and the...
most difficult items appear toward the end of the test. For the testing program studied here, the rationale for ordering items so that the most difficult items are in the middle of the exam is that students have warmed up by the middle of the test but are not yet fatigued, so that their concentration is thought to be at its most optimal. It should be noted that the tests in this program are untimed, so there are no concerns about test speededness. With timed tests, an additional rationale for

FIGURE 6  True minus estimated TCCs for the grade 5 math test by ability level with no item position change effects simulated (upper panel) and with item position change effects simulated (lower panel) for the 1000/5000 condition. Note: Plotted points represent the mean difference between the true and estimated TCC and lines extending from the points represent one standard deviation of these differences.
including the most difficult items at the end is to reduce the potential penalty on students who find themselves running out of time.

The effect of changes in item position on equating results would be much more severe if this program ordered items in the operational tests according to conventional practice, that is, from easiest to most difficult. To illustrate this effect, the simulations were repeated using an “easiest to most difficult” ordering.
based on the field-test item difficulties. To make the ordering realistic, a $N(0,1)$ random error component was added to the field-test $b$-values chosen at each iteration, and sum of the original field test $b$-value and the $N(0,1)$ error component was sorted from low to high to determine the final form item positions. Figure 9 summarizes the relationships between the field-test Rasch difficulties and the final form item positions for test forms selected in 10 simulation iterations using this
alternate approach. In this figure, the most difficult items tend to be placed toward the end of the test, although the relationship is not perfect due to the influence of the random error component.

Figure 10 summarizes the true minus estimated TCC differences over fixed theta levels for simulation sample sizes of 500 (field-test) and 2,500 (final form) over 100 replications where the operational items were approximately ordered from easiest to most difficult. These results indicate that when final form items are ordered in this manner, the changes in item position impact the equating results such that estimated scores at high ability levels are lower than the true scores, and vice versa. At the lower ability levels, the differences are approximately 0.3 of a raw score point and at the higher ability levels the differences are approximately 0.2 of a raw score point. These differences, although relatively small, could have a tangible impact on equating results. In addition, the direction of the change is systematic in that the easy items are getting easier and the more difficult items are getting more difficult.
DISCUSSION AND CONCLUSIONS

The results of this study indicated that although item position change from field testing to live testing can impact Rasch item difficulties, the observed effects for the testing program examined in this study were mitigated by the practice of ordering the field-test items such that easier items appear at the beginning and end of the test while more difficult items appear in the middle of the test. Although this ordering of items in the operational form was originally done for different reasons, it turns out to be very beneficial for the equating design employed with this testing program. That is, the simulations indicated that the effects on item difficulty due to position change essentially cancel each other out under this design.

However, the simulations in which final form items were ordered by putting the easier items (based on field-tested b-values) in the beginning of the test and the more difficult items toward the end of the test did indicate measurable effects related to test equating. These effects reflected the modeled relationship between field-test to final form item position change and b-value changes: difficult items became more difficult as they moved toward the end of the test, and easier items became easier as they moved toward the beginning of the test. As Figure 10 indicates, the resulting impact is that the test seems more difficult than it truly is for higher ability students and easier than it truly is for lower ability students. This would benefit higher ability students and disadvantage lower ability students in a resulting equating.
Clearly, these results justify the concerns expressed in the literature about item positions shifting from one administration to another when the IRT item parameter estimates for these items are utilized in the subsequent equating procedures. A strategy of field testing items together in the middle of the test seems to be tenable, although it seems to be somewhat dependent on the strategy used to order items in the final form. Another seemingly feasible strategy would involve field-testing items in an interspersed fashion throughout the test and placing limits on how much the position can change when the item is used in a final form (e.g., 5 positions or less). Such rules could work well for content areas involving discrete items. However, when reading passages are involved, it is much more difficult to restrict test construction so that items are ordered in a prescribed way or the number of positions an item can shift is limited. One flawed strategy of obtaining field-test item parameter estimates, assuming such estimates are to be used in subsequent equating procedures, is to place all field-test items at the end of the test. Based on the prediction equations obtained in this study, as well as the literature on item context effects, this practice would almost certainly make the operational test appear more difficult than it truly was. As a result, such a practice could give the false appearance of a sustained improvement in student performance over time.

This study was limited in several ways. First, we looked at only one large K–12 testing program that utilizes a particular measurement model, equating procedures, test administration procedures (i.e., an untimed test), and test construction procedures. These program-specific factors limit the generalization of findings to other testing programs. In addition, the simulations conducted to investigate the effects of the relationships between item position change and item difficulty change assumed all changes followed the polynomial regression model, and that the only other component affecting item difficulty was estimation error. Although these assumptions seemed reasonable for the purposes of the study, undoubtedly other factors exist that impact this relationship that we did not consider in our simulations.

In large-scale, high-stakes testing, which has increased exponentially under the No Child Left Behind legislation, state testing programs should continually evaluate their practices to ensure that they result in reliable results and valid inferences. Testing programs sometimes utilize item statistics obtained from field-testing to support the equating and scaling of operational tests. The results of this study illustrated that—in the context of one large-scale testing program—the effects of item position changes between field-testing and final form use can differ depending on multiple factors, including the original field-test positions and the way that field-test items are sequenced once they are used in final forms. Future research is needed to compare different strategies for embedding field-test items (e.g., located together in the middle of the test, placed in two different locations, interspersed throughout the test, located at the end of the test), and should investigate how these strategies interact with practices for sequencing the items in operational test forms. Finally, additional empirical data are needed to
confirm or contradict the models developed in this study to predict item difficulty change as a function of item position change, particularly from testing programs with characteristics that differ from the testing program utilized in this study.

ACKNOWLEDGMENT

The authors thank the reviewers for their extremely helpful comments and suggestions toward the improvement of this article.

REFERENCES


